Paweł Rutkowski, Joanna Filipczak, Anna Kuder, Piotr Mostowski
University of Warsaw, Section for Sign Linguistics

p.rutkowski@uw.edu.pl

Data acquisition, annotation and analysis in the corpus of Polish Sign Language (PJM)

Polish Sign Language (*polski język migowy*, usually abbreviated as PJM) is a natural visual-spatial language used by the Polish Deaf community. The aim of this paper is to present a large scale research project aimed at documenting PJM. Its main goal is to create an extensive and representative corpus of video material that will further form the basis of detailed grammatical and lexical analyses.

The underlying idea of the PJM corpus project is to record video clips showing Deaf people using PJM in a variety of different conversational contexts. When completed, the project will involve approximately 150 informants. The PJM corpus is diversified geographically and the group of signers participating in the project is well balanced in terms of age and gender. Data is collected exclusively from signers who either have deaf parents or have used PJM since early school age.

Recording sessions always involve two signers and a Deaf moderator. The procedure of data collection is based on an extensive list of tasks to be performed by the two informants. Typically, the signers are asked to react to certain visual stimuli, e.g. by describing a scene, naming an object, (re-)telling a story, or explaining something to their partner. The elicitation materials include pictures, videos, graphs, comic strips etc., with as little reference to written Polish as possible. All the necessary instructions are given in sign language exclusively; they have been pre-recorded and, like the elicitation materials, are presented to the participants on computer screens. The participants are also requested to discuss a number of topics pertaining to the Deaf. Additionally, they are given some time for free conversation (they are aware of being filmed but no specific task is assigned to them). The latter two parts of the recording session scenario are aimed at collecting spontaneous and naturalistic data.

When designing the above procedures, we took into account the challenges and problems encountered in similar projects conducted for other languages, in particular for German Sign Language (DGS) and Sign Language of the Netherlands (NGT). For instance, we attempted to make use of elicitation materials that had proved successful in the other projects.

The raw material obtained in the recording sessions is further tokenized, lemmatized, glossed and translated using the iLex software developed at the University of Hamburg. The annotation conventions we employ have been designed especially for the purposes of PJM (they are aimed at annotating both manual and non-manual signals).

The goal of the present paper is to give a detailed overview of the above procedures and show sample clips extracted from the PJM corpus in order to illustrate the most important advantages and disadvantages of the annotation choices that we have made.